LCRC Early User Update

Rémy Evard, Susan Coghlan, JP Navarro

16 January 2003

Planned Topics

- This is a free-form meeting, so please jump in at any point.

- Status
  - Usage and Experiences
  - Big Problems
  - Software Environment
  - Scheduling software and policies
  - Schedule
  - LCRC Hiring
  - Remaining major action items

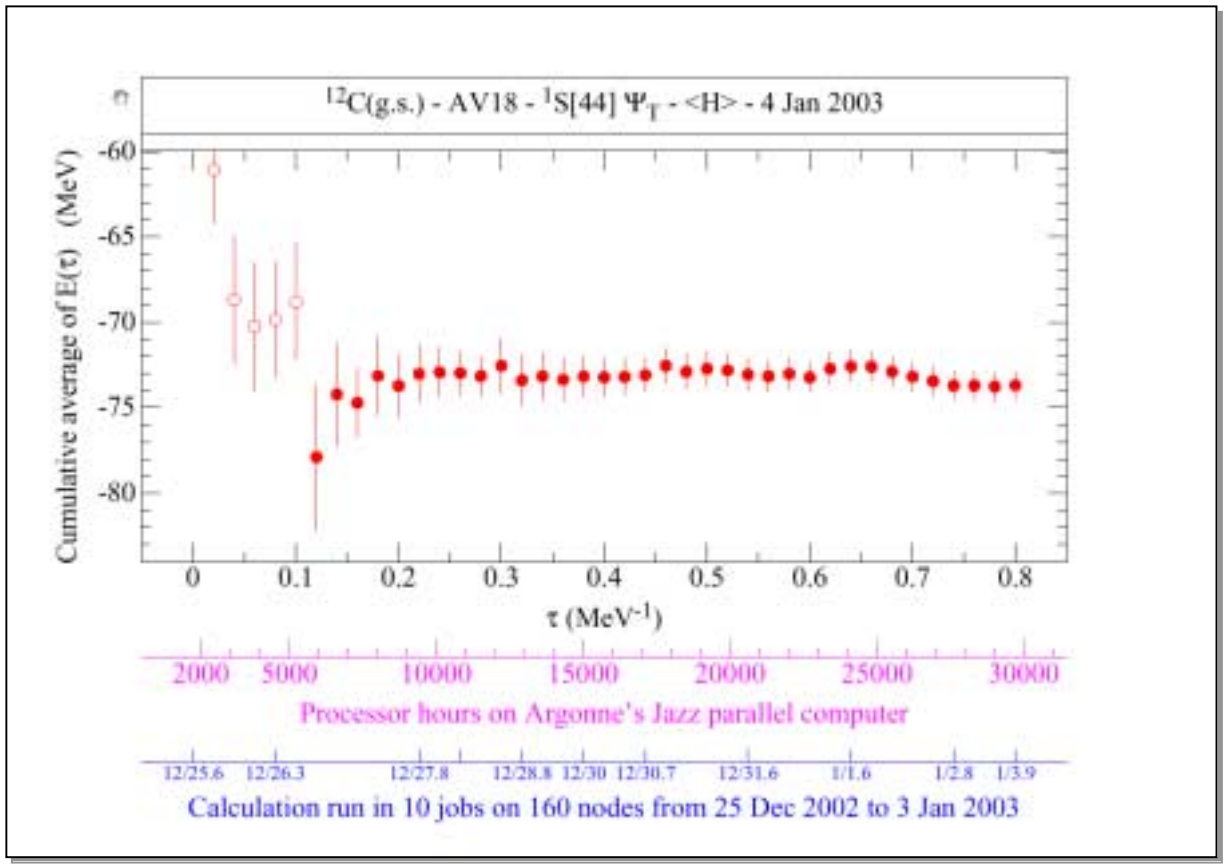
- Specific feedback questions

Status – Usage & Experiences Thus Far

- Early users
  - 32 early users (in addition to LCRC systems, vendors, etc)
  - Fields covered are diverse, and include neuroscience (3), chemistry (4), climate (2), computational fluids (3), bioinformatics (2), computer science (7), astrophysics (1), …
- Accounting information
  - Record-keeping began on November 15th.
    - Scheduler issues are causing some interesting data discrepances
    - Better record-keeping began in January…
  - Bulk of usage thus far has been:
    - LCRC systems staff and vendors for setup and debugging
    - Steve Pieper, PHY
    - Neuroscience – MCS/UofC
- Successes so far:
  - Performance appears to be quite good.
    - Robert Jacob, Rob Ross, and David Jones have all reported good performance.
  - Steve Pieper's results on next slides

Steve Pieper's results as presented in January PHY Colloquium

- Jazz processors 67% faster than NERSC Seaborg processors
- Biggest GFMC calculations they've done
  - 29,700 processor hours
  - 51.1 PFLOP
- Full calculation feasible?
- Geesaman said "state of the art"



Status – Big Problems

- Job submission and scheduling
  - Jobs immediately go into the blocked queue with deferred state
  - Status: actively working with Maui and PBS developers to solve this.
- NFS cache coherence
  - Status: creating instructions for users
- Intel 7 F90 Large MPI buffers
  - Large MPI buffers can be lost without users noticing it.
  - Status: Myricom working on this.
- MPI Iprobe bug – GM only
  - Myrinet jobs that have many 100,000's of MPI messages fill up memory allocation and die.
  - Status: Myricom working on this
- Large memory allocation
  - Unable to grab > 1GB contiguous memory
  - Status: exploring kernel fixes and other options
- MPICH Node Limit
  - No more than 256 nodes per job with MPICH rsh.
  - Status: shifting to mpd-based job launch
- xpvm sometimes kills nodes
  - The master node ethernet interface shuts down.
  - Status: Currently an unsolved bug.
- Intel MPICH can't find mpif.h
  - Status: Workarounds exist, will document this…….

- Notable things not on this list: no major GFS problems, no major hardware problems!

Status – Planned Software Environment * == in now

- Commercial Software purchased
  - ABSoft Compilers *
  - NAG Compilers and Libraries
  - Intel Compilers * and Libraries
  - Portland Group Compilers *
  - Totalview *
  - IDL
  - Matlab *
  - Gaussian/LINDA *
  - StarCD
- Requested
  - Insure++
    - (Need to figure out budget/licensing options.)
- Open/Public Software
  - Gnu compilers *
  - MPICH *
  - ROMIO *
  - MPD +
  - MPICH-2
  - MPD2
  - Globus
  - Columbus
  - NetCDF
  - NCO
  - NCAR
  - better blas
  - Accelrys
  - PETSc +
  - scalapack
  - PGA programming environment
  - Ccache *
  - python2 *
  - autoconf *
  - flex *
  - bison *
  - PVM *
  - XPVM *
  - X11 *
  - Emacs *
  - Bitkeeper *
  - CVS *
  - TCL/TK *
  - blas *
  - lapack *

Jazz Scheduling Software

- PBSPro 5.2.2 Resource Manager
  - Tracks job queue, nodes information, and starts/stops jobs
  - LCRC license with vendor support
  - Mostly stable (better than Chiba)
  - Vendor has already fixed one problem
- Maui 3.2.5 Scheduler
  - Decides when jobs run and on which nodes
  - Latest production code
  - Good developer support
  - Continuing to work on problems (mostly PBSPro interactions)
  - Expect ongoing debugging. Usage decision not made.
- Qbank 2.10.4 Allocations Manager
  - Tracks and enforces total node-hour allocation usage for projects and users
  - Latest production code
  - Good developer support
  - Working on integration w/ account/project request system
- Lots of commands!

Jazz Scheduling Policy

- Allocations per the LCRC allocation subcommittee
  - People will get initial startup allocation
  - Eventually will need a regular project allocation for continued use
  - May be enforced; credit / soft limit / ignore options
  - Users and PIs will be able to track balances and usage
- Job limits
  - We have the ability but no existing plans to enforce fair-share limits: number of nodes, duration, nodes * duration, number of queued jobs, number of running jobs, etc.
  - Considering a weekday “X node-turnover” policy
  - Considering a shared node job partition (different charge rate)
  - Scheduled downtime

- We need your feedback so we can send reasonable recommendations to CSAC for final decision.

Status – Update on LCRC personnel

- LCRC Scientific Application Engineers (Scientific Consultant)
  - 2 positions are now open
  - Several internal people have applied, interviews begin next week
  - Also looking for external applicants
  - Will send email with the URL for the job info

- LCRC full-time systems staff:
  - Susan Coghlan
  - Rick Bradshaw

- LCRC partial assistance
  - John-Paul Navarro (HPC services), Gene Rackow (facilities, OS, security), Sandra Bittner (HPC software), Caren Litvanyi (networking), Narayan Desai (early configs).
  - As the facility stabilizes, much of this work will get focused and we'll determine whether or not to hire a dedicated LCRC staff member to pick it up.

Status – Remaining Major Action Items

- Fixing the Big Problems
- Continued software installs
- Continued response to issues sent to systems@lcrc.anl.gov

- Web-based account requests
  - Nearly complete
- Web-based project management
  - Be able to request allocations via the web, assign project percentages to project participants, etc.
  - Nearly complete

- LCRC Web pages
  - Current system information, tutorial, FAQs, etc
  - Just starting these pages

Status – Schedule

- January
  - Major action items
  - Hiring moving forward
  - LCRC allocation policy subcommittee decisions
  - Continue to add accounts for any who request them
- February
  - Go live with project accounting and scheduling policies for early users
  - Go live with web-based accounts for new users
  - Announce availability to entire Lab

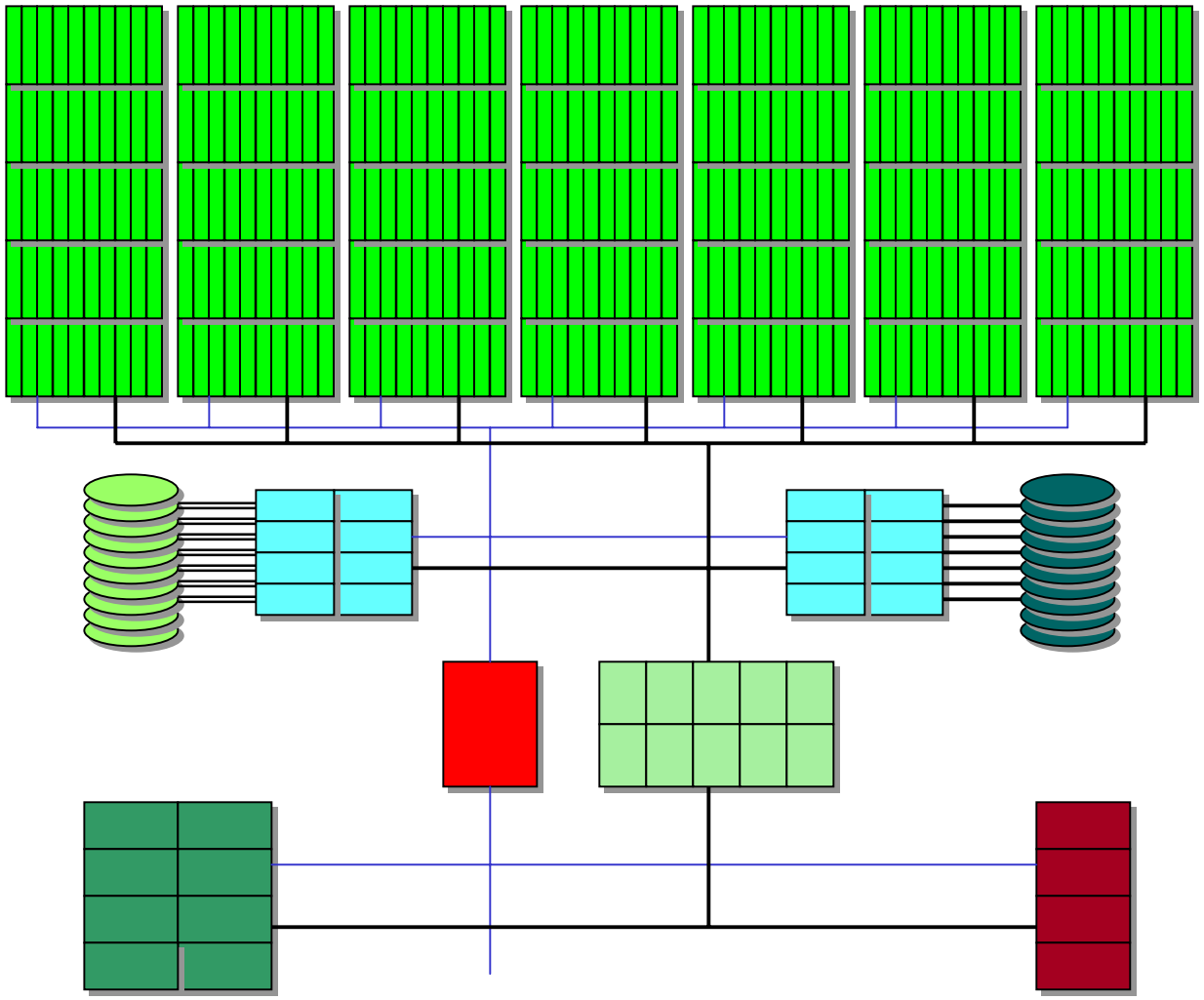
Feedback

- Are there any configuration changes that should be made?
- Are there any problems that we're not aware of?
- We're working on a tutorial for new users. Are there any specific things that you recommend should be in it?
- Announcements about Jazz activity – how would you prefer to get these?
  - Email?
  - When you login?
  - More or less than current?
- Questions?

- In general, please send email to **systems@lcrc.anl.gov**

Jazz - the ANL LCRC Computing Cluster

Current Configuration, January 2003



1Gb to ANL

350 computing nodes:

2.4 GHz Pentium IV
50% w/ 2 GB RAM
50% w/ 1 GB RAM
80 GB local scratch disk
Linux

10 TB global working disk:

8 dual 2.4 GHz Pentium IV servers
10 TB SCSI JBOD disks
PVFS file system

10 TB home disk:

8 dual 2.4 GHz Pentium IV servers
10 TB Fiber Channel disks
GFS between servers
NFS to the nodes

Network:

Myrinet 2000 to all systems
Fast Ethernet to the nodes
GigE aggregation

Support:

4 front end nodes: 2x 2.4 GHz PIV
8 management systems